BIG DATA PROCESSING AND ANALYSIS ON THE IMPACT OF COVID-19 ON PUBLIC TRANSPORT DELAY

Yuming Ou, Adriana-Simona Mihaita, Fang Chen

Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

yuming.ou@uts.edu.au, adriana-simona.mihaita@uts.edu.au, fang.chen@uts.edu.au

ABSTRACT

The COVID-19 pandemic that started at the beginning of the year 2020 has significantly disrupted people's daily life around the world. Understanding and quantifying the impact of such a large-scale disruption will help people mitigate the pandemic and enhance the resilience for future preparation of similar events. In this chapter, we present a research work studying the impact of COVID-19 on public transport in terms of bus delay, which involves big data processing and analysis on multi-source data sets containing COVID-19 case data, bus GTFS (General Transit Feed Specification) data and LGA (Local Government Area) boundaries data. The datasets in use are heterogeneous, arrive in large-volumes and in real-time, have a spatial-temporal distribution, which brings true challenges to this research. To quantify the bus delays changes, we propose a methodology consisting of real-time data crawling, map-matching, arrival time estimation, and bus delay calculation and aggregation. The methodology is applied to a case study focusing on Sydney metropolitan region across different stages of COVID-19 pandemic from February to March 2020. The case study shows that during March 2020, the COVID-19 pandemic has significantly impacted people's travel behaviours in Sydney, though the influence varies in different areas. The most affected areas are the central and eastern suburbs which recorded a drop of 9.5 minutes of bus delay during afternoon peak hours. The findings are helpful to understand and mitigate the restriction impact in different city areas with different conditions. The quantified delay reduction also reveals the potential of better transport performance which could be used as a benchmark of transport performance improvement after the pandemic. The main contributions of this work include the methodology to quantify travel behaviour changes under large disruptions such as COVID-19 pandemic and the case study on large-scale and long-period travel behaviour shift that seldom happened before.

KEYWORDS

Big data analytics, Public transport, Bus delay, COVID-19 pandemic, GTFS data

1. INTRODUCTION

The recent COVID-19 pandemic triggered at the beginning of the year 2020 [1] has significantly affected people's daily life around the world from all aspects: personal transport mobility (both local, national and international), daily working schedule, tourism activities, medical operating conditions, etc. Understanding and quantifying the overall impact of such a large-scale disruption will help citizens to mitigate the pandemic and enhance their resilience for future preparation of similar events. To this end, in this chapter, we present the research work that we conducted to study the impact of COVID-19 on public transport at different stages in different areas across a large metropolitan area from Sydney, Australia, particularly in terms of bus delay. The public transport infrastructure in Sydney is very important as it connects radially the Inner and East region of the city to the large extended regional suburbs. As the city does not have an underground subway system like other large metropolitan cities from Europe or U.S.A, the bus public transport system is critical in delivering the daily home-to-work journeys. The results not only help to understand the changes of travel behaviours caused by the pandemic across the entire city, but also to benefit the public transport operators to enhance services after the pandemic will be finished, by providing a benchmark of potential improvement.

In Australia, a few cases of COVID-19 were initially reported before March 2020. However, the number of confirmed cases has rapidly increased during the month of March. There was a signalised spike of reported cases during 22nd of March 2020 to 27th of March 2020 in the state of New South Wales (NSW). Based on this observation, we hypothesize that under the situation of rapid increase of confirmed cases, people will largely reduce their usage of private cars and public transport by working from home and avoiding unnecessary travel. The decision of reducing daily trips was also recommended by government authorities to keep social distancing and work from home where possible [2]. The travel behaviour shift has led to a change of traffic condition, which had no precedent. Therefore, the impact of such travel restriction can be quantified in terms of bus delay, which is an important indicator of the overall urban traffic condition. With this hypothesis, we decide to scope our study on the bus delay change in the Sydney region within the time window from February to March 2020. As an observation, previous years of bus delays analysis have been captured continuously by the research team, but for the purpose of this work and case study, we keep the data analysis restricted around the most affected months by the global COVID-19 pandemic. We do make the observation that while a great importance has been previously given in the literature to the prediction of bus delays in the network, the current large scale disruption of COVID-19 represents a unique event which hasn't been studied before, therefore the challenge to model and understand the impact of such disruption on not only bus operations, but also on the traveller's behaviour.

Besides the difference of bus delay between February and March, we also aim to study the public transport differences recorded around different urban areas, in order to have a more comprehensive understanding of the COVID-19 impact on urban mobility. In Sydney, the CBD and eastern suburbs are the very popular locations for business, travelling, transport multi-modal hubs and tourism attractions due to their proximity to beaches. This makes the travel behaviour in these areas distinct from other areas in the city which have different economical and urban characteristics (several warehouses, depots, remote industrial headquarters, local suburban houses). Consequently, the impact of COVID-19 on these areas is expected to be different from the remote areas, especially during peak hours.

In order to quantify the bus delay changes, the biggest challenge is to correctly and timely estimate the bus arrival times at various bus stops in the network. The information of bus arrival time is currently unavailable as the bus stations are not interconnected or monitored digitally. However, the real-time positioning of each bus across the network is available via real-time transmitted GPS (Global Positioning System) location points. We therefore need to estimate the bus arrival times by using the real-time transmitted GPS data, which contains the geolocations of buses in the city with their associated time stamps. Nowadays the GPS devices mounted on buses are continuously generating data within a very short time-interval of few seconds, with the purpose to better monitor bus movements. This makes the accurate estimation of bus arrival time possible. On the other hand, due to the extensive and large bus network, this requires having in place a big-data processing and analysis framework to run on such a large-volume of geolocation data set; for example, more than 3 GB of GPS data is generated every day for the entire Sydney bus network. Furthermore, due to the well-known issue of GPS accuracy [3], the GPS data is always associated with an error (a deviation from what the real position of the bus vehicle is in reality). Therefore, we are facing the challenge of mapping the GPS data points to the correct geolocations in order to be able to calculate accurately the individual bus delays, and aggregate results for the impact analysis.

This chapter is organised as follows. In Section 2, we explain the data sets used for this research and the methods to collect the data, after which we propose a methodological framework of data processing and analysis to quantify bus delays in Sections 3. Section 4 reports a case study on Sydney metropolitan region across different stages of COVID-19 pandemic. Finally, we conclude this chapter and provide some directions for future work in Section 5.

2. DATA PREPARATION

The data used in this research work is heterogeneous and comes from multiple sources, including COVID-19 case data, bus GTFS (General Transit Feed Specification) data and LGA (Local Government Area) boundaries data. As shown in Table 1, the data that we are using has several characteristics: it has a temporal-spatial distribution (data points are transmitted with high time frequency across a very large area coverage) and arrives in real-time and in a large-volume; this brings various challenges to the data processing and analysis. In the following, we detail the above three types of data and their corresponding collection methods.

Table 1. Data sets used in this research work

Data	Temporal	Spatial	Real-time	Large volume
COVID-19 data	Yes	Yes		
Bus GTFS data	Yes	Yes	Yes	Yes
LGA boundaries data		Yes		

2.1. COVID-19 Case Data

COVID-19 data contains the information for all confirmed case in NSW Australia including the notification date, the LGA code (where it happened in the city) and the LGA name. The data is available on the official government website Data.NSW [4] and is updated on a daily basis. A sample of the COVID-19 data is shown in Table 2.

Table 2. Sample of COVID-19 data

Notification Date	LGA Code	LGA Name
2020-01-22	11300	Burwood (A)
2020-01-24	16260	Parramatta (C)
2020-01-25	14500	Ku-ring-gai (A)
2020-01-25	16550	Randwick

To understand the trend of increased confirmed cases, we group the data by their notification date and visualise the results in Figure 1. As shown in Figure 1, there were 10 confirmed cases before March 2020. Starting with 1st of March 2020, the number of confirmed cases was rapidly increasing and reaching almost 200 cases daily. The peak occurred during the time window from 22nd of March 2020 to 27th of March 2020 while slowly decreasing afterwards.

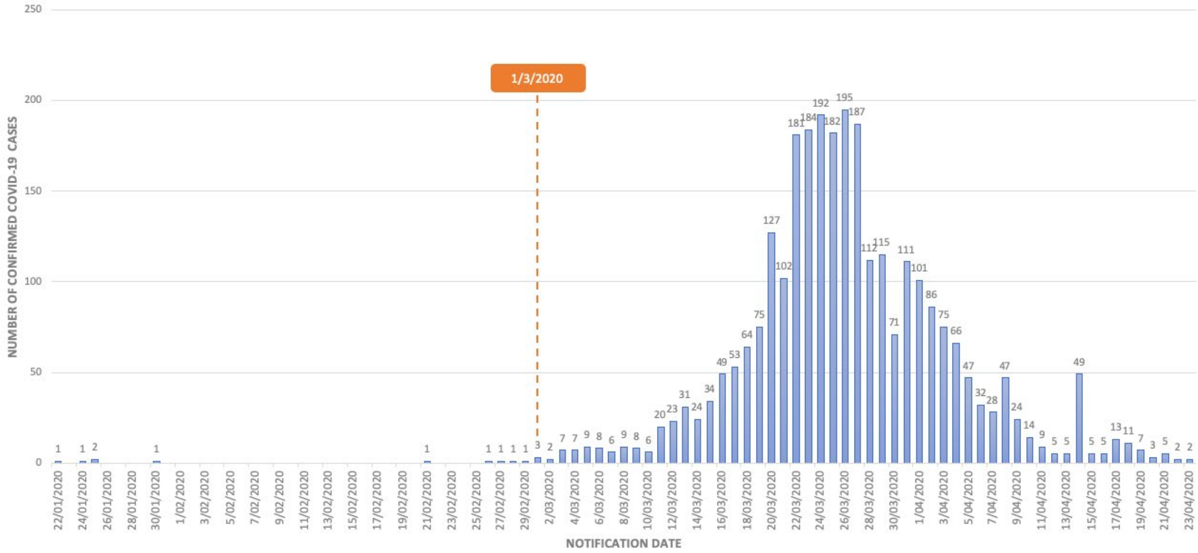


Figure 1. Epidemiological curve of confirmed COVID-19 cases by notification date, NSW, Australia.

2.2. Bus GTFS Data

GTFS [5] is the abbreviation for General Transit Feed Specification, which defines a common data format to allow public transit agencies to publish their transit data so that the data can be consumed by various applications. Generally, GTFS is divided into GTFS static and GTFS real-time streams. The former contains public transportation schedules and associated geographic information while the latter contains the real-time vehicle positions and all trip updates.

Nowadays GTFS has been used as an industry standard for majority of transit agencies to publish their transit data around the world [6]. As GTFS data contains both scheduled and real-time information about transit operations, it has been actively used for many research problems such as transit accessibility [7, 8, 9, 10, 11], transit network analysis [12, 13], performance evaluation [14, 15], delay prediction [16, 17, 18], and transit trip inference [19, 20].

In this research, we use the following three types of GTFS data:

- *Real-time bus position data*: the real-time buses' movements with longitudes, latitudes, and associated time stamps. The real-time bus positions are captured by the GPS devices mounted on the buses. As mentioned before, there are always errors associated with the GPS data. We need to correct the GPS data by a map matching algorithm which we introduce in Section 3 of this chapter. Figure 2 shows an example of how such GPS point coordinates can be transmitted with some of them deviating from the road centreline.

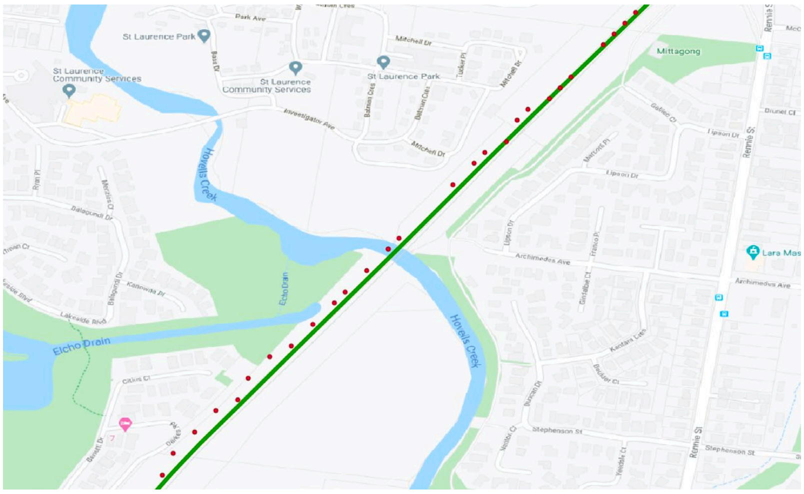
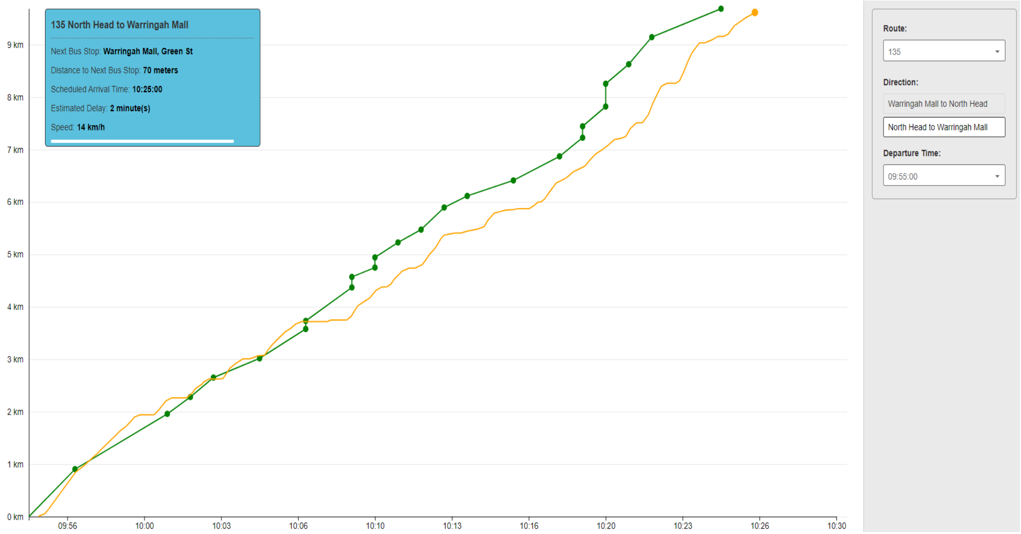


Figure 2. Example of GPS point corrdinats transmitted across a transit line.

- *Bus timetable data*: contains the scheduled bus trips and scheduled arrival times at bus stops. Figure 3 showcases an example of how a bus timetable is being captured in the urban region; the *x* axis represents the timeline from the departing time of the bus trips 135 (in this case scheduled to depart at 09:55 AM) and until the end of the bus stop (10:26 AM), while the *y* axis represents the distance travelled since the departure of the bus. The green line showcases the planned timetable trips, while the orange line represents the real-tracked movement of the bus while in operation. One can notice that the bus in this example was always on time and even arrived earlier until 10:06 AM, after which it started to record accumulated delay until the end of the trip. Similarly, we conduct the same visualisation and tracking for all bus trips across the entire network. The difference between the scheduled versus the real-tracked movement of the bus at each stop is identified as the “bus stop delay”.

- *Bus network data*: which contains the geolocations of all bus stops (as represented in Figure 4) and the physical geometry (layout) of the bus routes (as represented in Figure 5).



135 North Head to Warringah Mall

Next Bus Stop: **Warringah Mall, Green St**

Distance to Next Bus Stop: **70 meters**

Scheduled Arrival Time: **10:25:00**

Estimated Delay: **2 minute(s)**

Speed: **14 km/h**

Route: 135

Direction: Warringah Mall to North Head / North Head to Warringah Mall

Departure Time: 09:55:00

Figure 3. Example of bus timetable data for a bus line (135) in Sydney, scheduled to depart at 09:55AM.

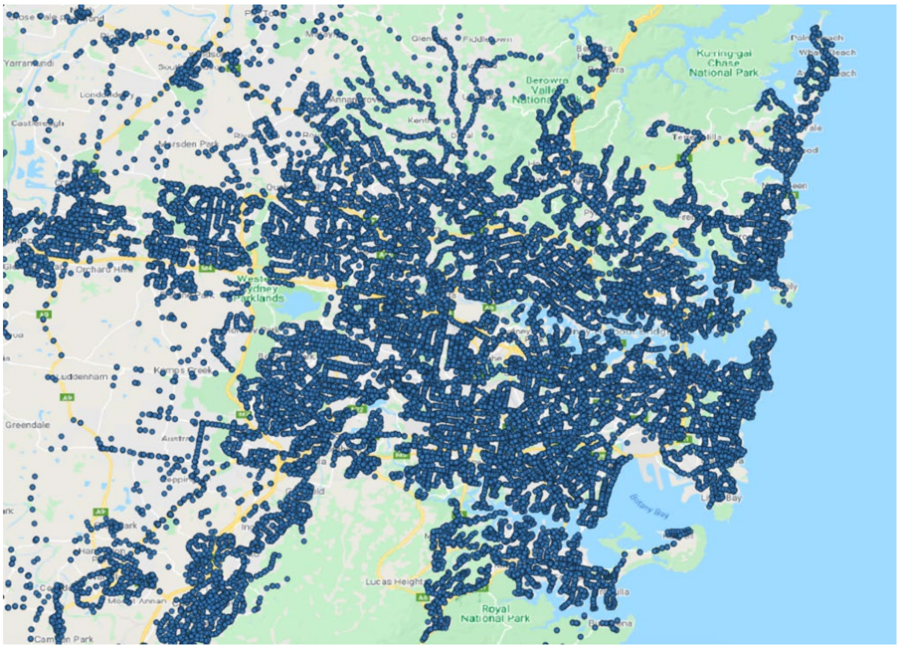


Figure 4. Geolocations of bus stops in Sydney metropolitan region. Each blue dot is a bus stop.

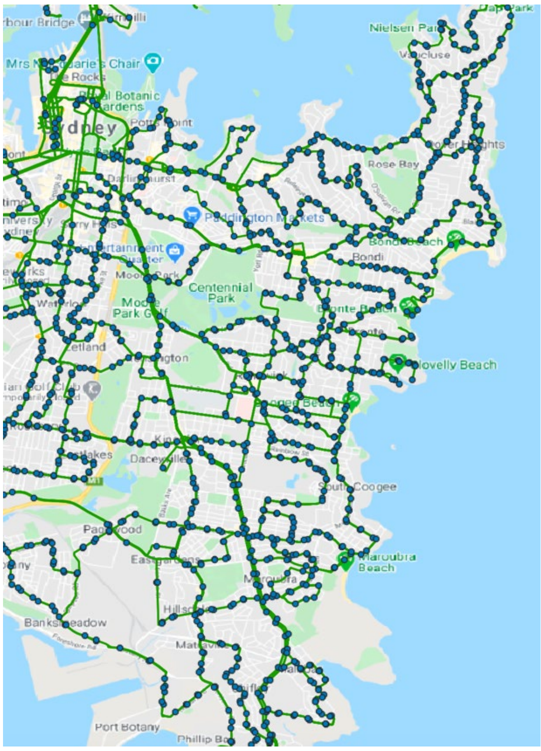


Figure 5. Bus routes and stops in Sydney CBD and eastern suburbs.

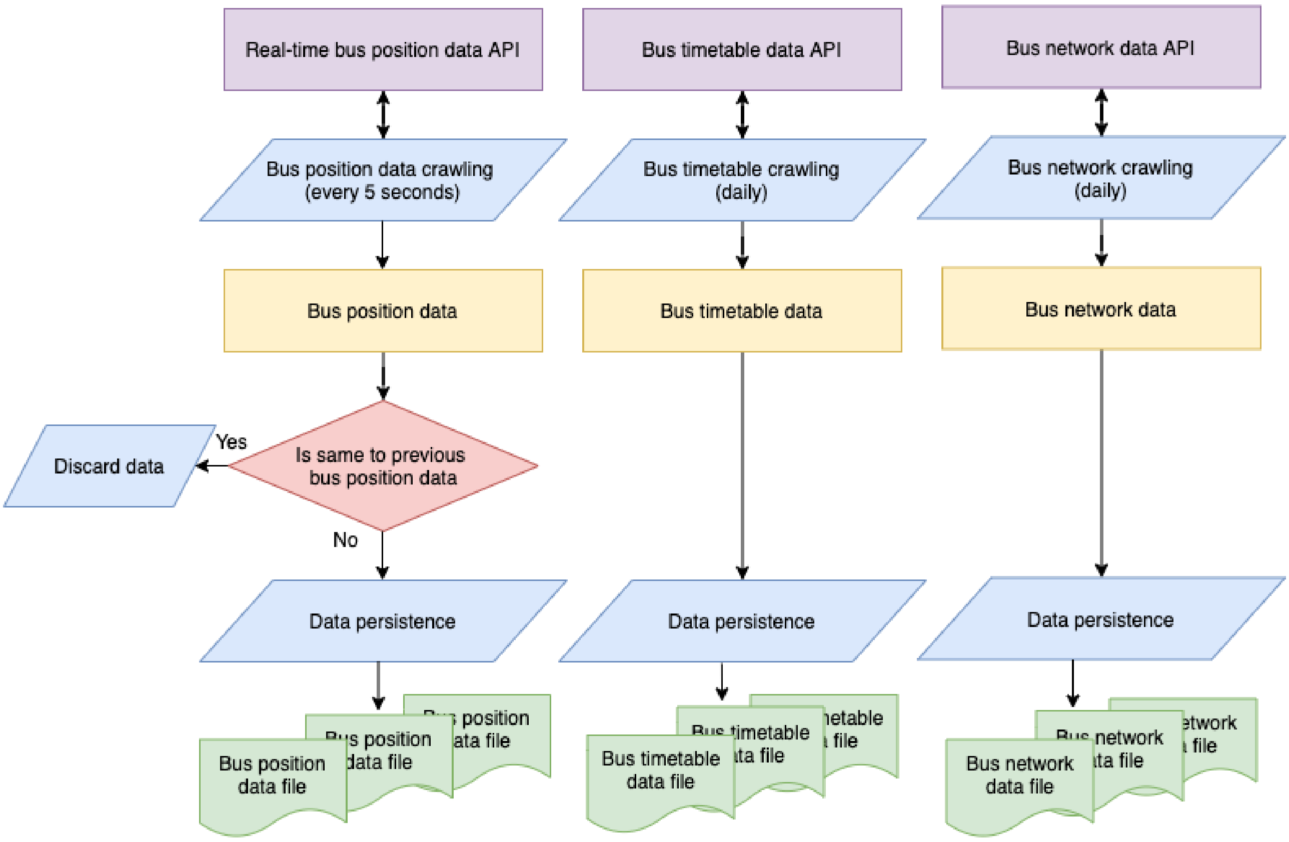


Figure 6. Workflow of crawling bus GTFS data.

As the bus GTFS data is published in real-time, we need to develop a data crawling service which is continuously collecting the data through RESTful data APIs provided by the local transport agency: Transport for NSW [21]. Figure 6 illustrates the workflow of data crawling. To collect the real-time bus position data, the data crawling service sends a data pulling request to the data APIs every 5 seconds. After receiving the data returned from the data API the service then parses the data and checks whether it is exactly the same to the previous data points by examining the hash values. If so then it discards the data, otherwise we append the data to stored data files. The purpose of removing the duplicate records is to save space as well as to reduce the computation cost in the step of delay calculation introduced in Section 3. In the entire Sydney metropolitan region there are around 24,000 bus stops and more than 25,000 bus trips are being scheduled during a 24-hour day, which leads to more than 3GB of real-time bus position data being collected every day. Apart from the bus position data, the data crawling service also collects timetable and network data daily in a similar fashion, in order to have up-to-date timetable and network data.

2.3. LGA Boundaries Data

The LGA boundaries data is downloaded from Australian Bureau of Statistic website [22], containing the information of LGA geometries and area catchment, which plays the important role to link COVID-

19 confirmed cases data to bus data in terms of geolocation. The LGAs in Sydney metropolitan region is shown in Figure 7.

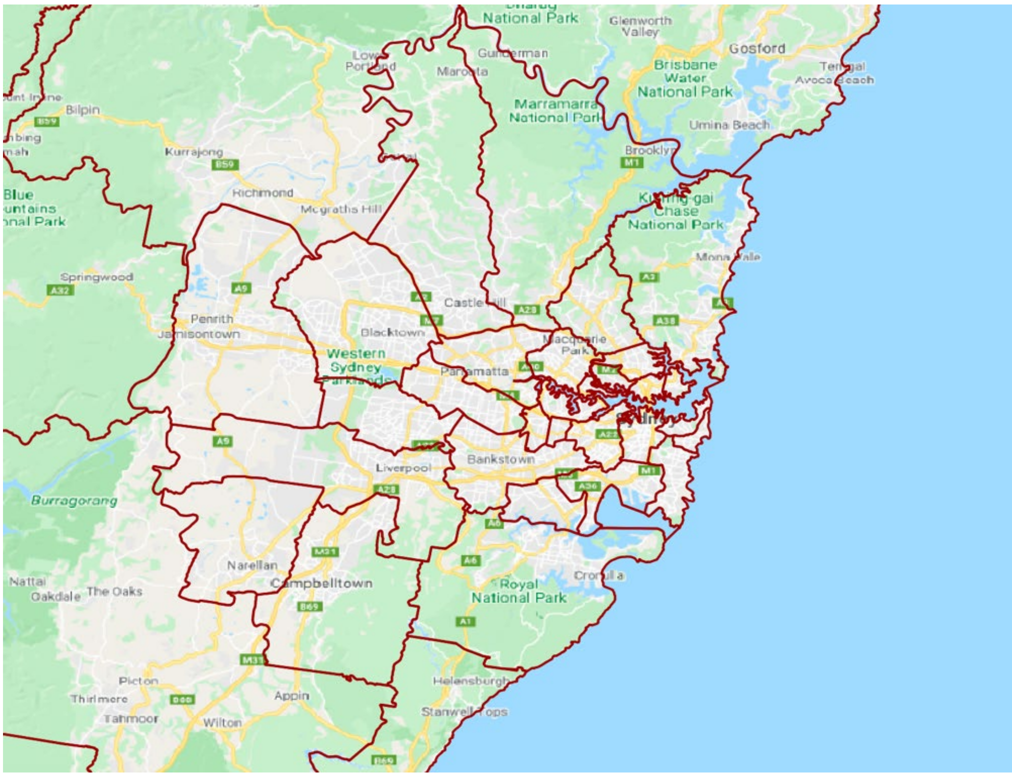


Figure 7. LGAs in Sydney metropolitan region. The read lines are the boundaries of LGAs.

3. METHODOLOGY

In this section we propose our methodology to quantify bus delays based on the collected data described in Section 2. We first introduce the framework and explain the workflow of each step, and then provide the details of each step.

3.1. Framework

Figure 8 illustrates our approach to quantify the bus delays, in which every blue block represents a main algorithmically processing step, every green block is a raw data set which was collected, and each orange block is the output of each processing step. They are explained as follows:

1) *Bus GTFS data crawling*. As introduced in Section 2, a data crawling service is continuously collecting bus GTFS data through data APIs and appending the data to files. The data collected includes bus position data, bus timetable data, and bus network data as previously described. The data crawling service is running as a standalone process and is independent to other steps.
2) *Map matching*. For each bus trip that is being collected, bus GPS points are matched to the actual geolocations in the bus network by using our proposed map matching algorithm. This step represents a true research challenge as the GPS points' deviation can be quite high

especially in central urban areas where various complex intersections can be found. Matching the bus GPS locations not only needs to consider the proximity to the road centreline, but more specifically, the direction of the bus trips, the turnings available at each intersection, the last GPS points, the overlaying with other bus trips sharing the same lanes, etc. The method is further described in subsection 3.2 of this chapter. The output of this step is the corrected bus geolocations with associated time stamps.

3) *Bus stop extraction*. This step is needed in order to extract the bus stop geolocations from bus network data collected by the data crawling service. The extracted bus stop geolocations are used in the step of arrival time estimation (step 5).
4) *Scheduled arrival time extraction*. This step is further applied in order to extract the scheduled arrival time at each bus stop for each bus trip from the bus timetable data collected by the data crawling service. The extracted scheduled arrival times are used in the step of delay calculation (step 6).
5) *Arrival time estimation*. Based on the bus geolocations with time stamps obtained in step 2 and the bus stops geolocations obtained in step 3, we track the bus movements and estimate the actual arrival times at each bus stops. The method to estimate the arrival time is detailed in subsection 3.3.
6) *Delay calculation*. By checking the arrival times against the scheduled arrival times, we calculate the delays at bus stops as the difference between planned and actual arrival time. The idea of delay calculation is initially introduced in Figure 3 and we provide the formalised calculation in subsection 3.4.
7) *Delay aggregation*. Finally, the delays at every bus stop for all trips are aggregated to each LGA level based on their geolocations and time stamps' time windows, utilising the LGA boundaries data. The aggregation method is detailed in subsection 3.4.

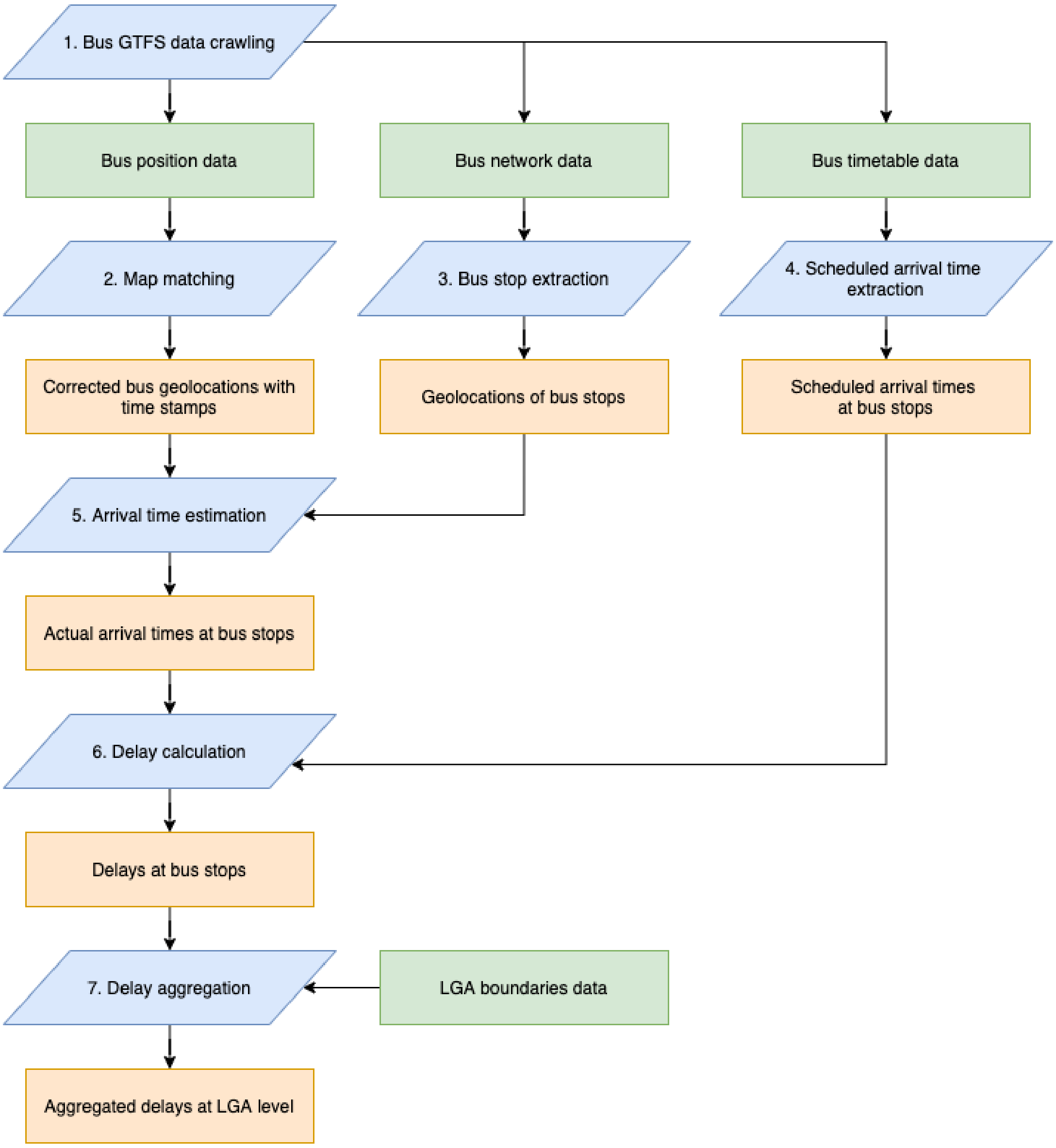

1. Bus GTFS data crawling
Bus position data
Bus network data
Bus timetable data
2. Map matching
3. Bus stop extraction
4. Scheduled arrival time extraction
Corrected bus geolocations with time stamps
Geolocations of bus stops
Scheduled arrival times at bus stops
5. Arrival time estimation
Actual arrival times at bus stops
6. Delay calculation
Delays at bus stops
7. Delay aggregation
LGA boundaries data
Aggregated delays at LGA level

Figure 8. The methodological framework to quantify bus delays at each zoning areas in the city.

3.2. Map Matching

As mentioned before, GPS data has an inevitable error which is variable depending on the circumstances, the road network geometry layout and continuity of data transmission in real-time. Many other sources could contribute to GPS errors, such as clock error, signal jamming, weather and building blocking. An example of GPS errors is shown is Figure 9 in which the red dots are the GPS data points sent from the GPS device on a bus while the green line is the actual bus trajectory along the main road. It can be observed that many GPS locations are falling further away from the green line (road centre line) instead of exactly being on it. Consequently, before using the bus GPS data to estimate arrival times, we need to correct the GPS data through map matching algorithms by matching every GPS coordinate transmitted by the bus to a correct location on the road centreline.

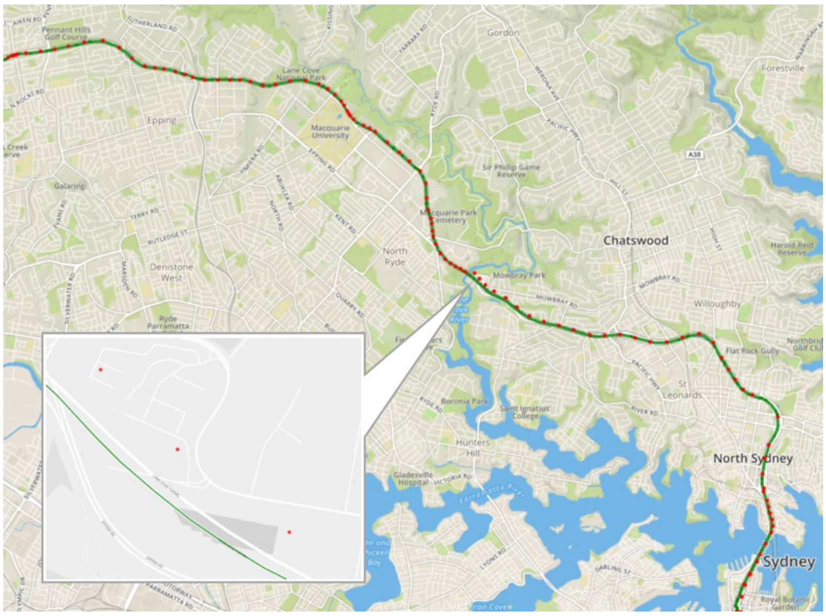


Figure 9. An example of GPS errors.

There are various methods that have been used in the literature for map matching [23, 24, 25, 26]. One native way is the point-to-curve method, which projects GPS points to their closest edges. This method is simplistic and lacks robustness especially when the road network has a complicated structure such as inside the CBD. An improved method is the curve-to-curve method which considers the closeness and similarity between the curve formed by GPS points and the candidate path. However, it still has the same problems under the circumstances of large GPS errors and complicated overlayed networks. Other approaches include using the geometry and topology of the road network [27], Kalman filters [28], and Fuzzy rules [29]. Generally, map matching algorithms can be divided into two groups: offline and online algorithms. Offline algorithms have an advantage of processing all GPS points after a trip is finished. On the other hand, online algorithms need to process currently available GPS points before a given time, which makes them have less data to be used and potentially leading to a compromise of accuracy.

To achieve a high accuracy of GPS data correction in real-time, our map matching method is based on a Hidden Markov Model (HMM) [30, 31]. HMMs usually model a system by considering their unobserved states and their observations. In the system one hidden state can change to any other hidden state by following a state transition probability. Instead of the hidden states, one can observe the values generated from the hidden states with emission probabilities. In this work, we model the road segments on which the bus is as the hidden states and the GPS readings as the observations as shown in Figure 10. Under this setting, the emission probability is defined in the following Equation 1,

$$P\left(GPS_t \middle| Seg_t^i\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{gd(GPS_t, GPS_t^i)^2}{2\sigma^2}} \tag{1}$$

in which GPS_t is the bus GPS reading at time t, Seg_t^i is the road segment *i* that the bus is on at time t, GPS_t^i is the projection of GPS_t on Seg_t^i, gd is the great circle distance between two geolocations, and σ is the stand deviation of the GPS error.

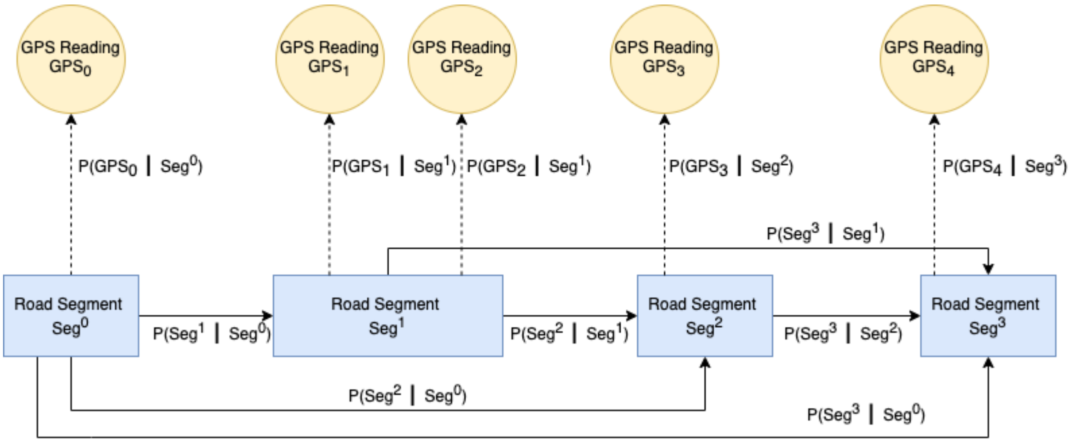


Figure 10. Hidden Markov Model for map matching.

Furthermore, the transition probability is defined in the following Equation 2,

$$P\left(Seg_{t+1}^j \middle| Seg_t^i\right) = \frac{gd(GPS_t,\ GPS_{t+1})}{rd(GPS_t^i, GPS_{t+1}^j)} \tag{2}$$

in which rd is the distance between two geolocations along the segment path.

Given a sequence of GPS readings as the observations, we can utilize the Viterbi algorithm [32] to find out the most likely sequence of road segments as the hidden states.

3.3 Arrival Time Estimation and Delay Calculation

After obtaining the actual bus geolocations, we then use them together with bus stop geolocations to estimate the arrival times at bus stops. The arrival time A_s^t at the bus stop *s* for bus trip *t* is computed using the following Equation 3,

$$A_s^t = \frac{rd(GPS_n^i,\ GEO_s)}{rd(GPS_n^i, GPS_m^j)}\left(T_{GPS_n} - T_{GPS_m}\right) \tag{3}$$

in which GEO_s is the geolocation of bus stop *s*, T_{GPS_m} is the time stamp of the last GPS reading GPS_m before bus stop *s*, $T_{GPS_n}$is the time stamp of the first GPS reading GPS_n after the bus stop *s*.

3.4 Delay Calculation and Aggregation

The delay D_s^t at bus stop s for bus trip *t* is calculated using the following Equation 4,

$$D_s^t = A_s^t - SA_s^t \tag{4}$$

in which SA_s^t is the scheduled arrival time at bus stop s for bus trip t. A negative D_s^t means that the bus will arrive at the stop early than the scheduled time, while a positive D_s^t indicates that the bus will be late.

Finally, after all delays have been calculated for all bus stops and all bus trips, we use LGA boundaries data to fuse the COVID-19 case data with bus delay data. By checking whether the bus stop geolocations are within a same LGA boundary, we divide the bus stops into groups in which all bus stops are located in a same LGA. We further calculate the median, the 5th percentile and 95th percentile for each LGA group and each 30-minute time window from 5:00AM to 22:00PM. We also group the number of confirmed cases of COVID-19 by their locations of LGA. In this way, the COVID-19 case data and bus delay data are linking to each other.

4. CASE STUDY

In order to verify our methodology, we conducted a case study to quantify the impact of COVID-19 in terms of bus delay in the Sydney metropolitan region. As previously mentioned, the time window of the case study is from February to March 2020 so all results presented here will be referring to these two months of study.

4.1. Area Selection

We first started with data processing on the COVID-19 records to visualise the accumulated number of confirmed cases in March for each LGA. Figure 11 (a) presents the results of accumulated infected cases by each LGA; the higher the number of reported cases, the darker the LGA blue area becomes. There were 6 LGAs with an accumulated number of cases reaching more than 80, which are grouped into four areas in terms of geolocation including:

- CBD and eastern suburbs,
- Northern Beaches area,
- the Southerland shire area, and
- the Blacktown area.

The area around city CBD and eastern suburbs ranks at the top of accumulated cases in March 2020 and it represents a higher priority for government agencies due to the importance and location of all major business hubs, together with all services and multi-modal transport hubs.

Following the methodology proposed in Section 3, we further processed the bus data and quantified the variation of bus delay from February to March for each LGA. By "variation", we refer to the reduction (or increase) in the aggregated bus delay for each LGA. The results are visualised in Figure 11 (b). The darker the colour of each LGA, the bigger was the delay reduction registered for that particular LGA.

Upon a detailed investigation, it can be observed that there were significant changes in three areas including:

- CBD and eastern suburbs,
- Northern Beaches area, and
- the Southerland shire area.

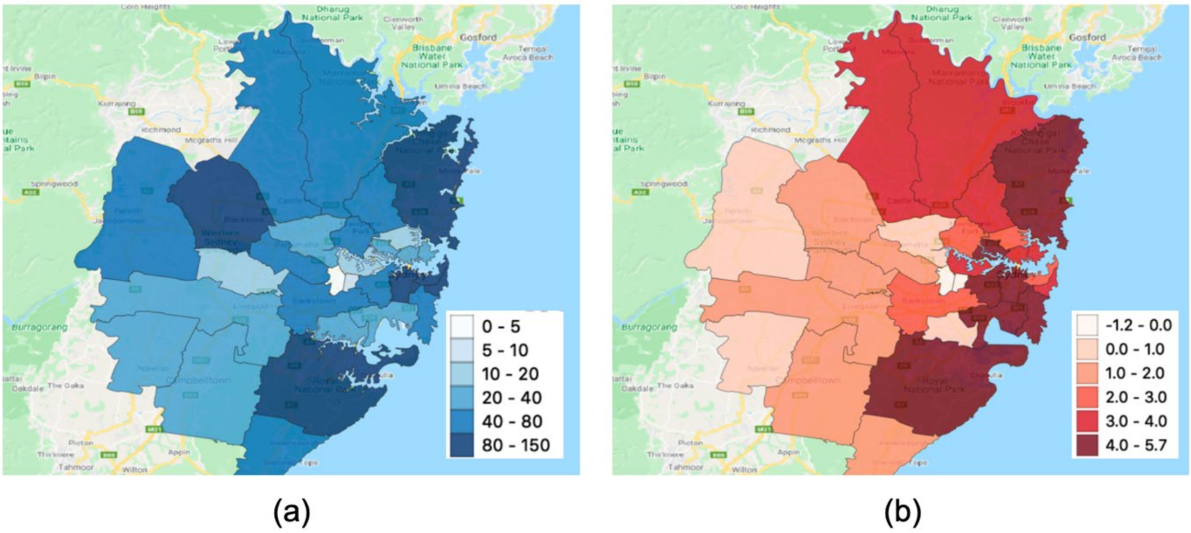


Figure 11. (a) Accumulated number of confirmed cases of COVID-19 in March 2020. (b) Change of bus delays from February to March 2020 (the darker the area, the bigger the delay was impacted).

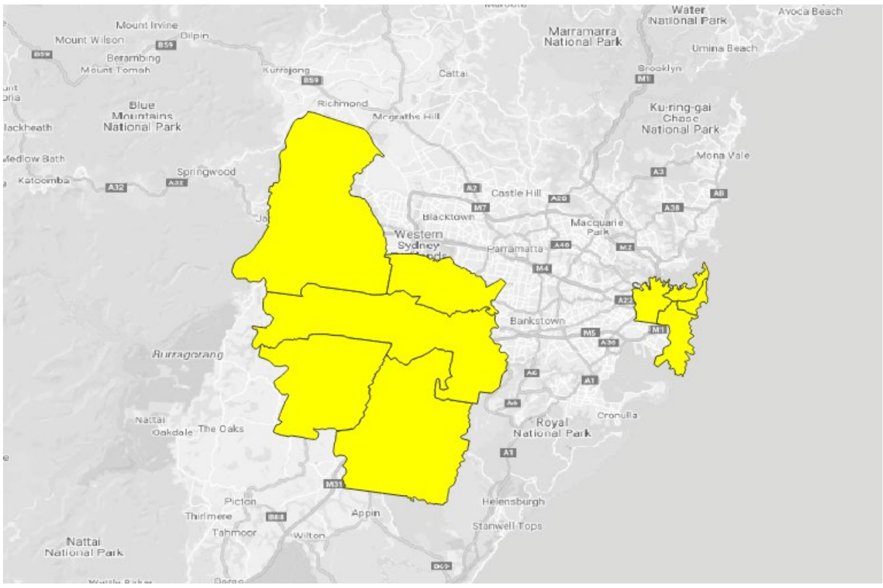


Figure 12. Area A and B selected for comparing the changes of bus delay.

It is notable that the area of CBD and eastern suburbs ranks as the top of the most affected area by a significant change in bus delay but as well in the number of accumulated infection cases. These interesting findings motivate us to further investigate the understanding of the impact that the COVID-19 pandemic has had on different urban areas. We strategically choose two areas with different characteristics for comparison. One area (called area A) is the area of CBD and eastern suburbs due to the above interesting findings. Another (called area B) is the area of Western suburbs which has a small number of confirmed COVID-19 cases. These two selected areas are shown in Figure 12.

4.2. Bus Delay Analytics and Comparison

Figure 13 and Figure 14 show the bus delay statistics on areas A and B in February and March respectively, including: the median bus delay for each area, the 5th and the 95th percentiles of bus delay from 5:00AM to 22:00PM, recorded across all bus stops in the area. Blue lines stand for the delays in February 2020 and orange lines stand for the delays in March 2020. It can be observed that the change of the bus delay was more significant in area A than in area B, especially in the upper bound of delay in peak hours. The overall average decrease of delay was 4.4 minutes (36%) in area A while only 1.3 minutes (15%) in area B in peak hours (7:00 AM - 9:00 AM and 16:00 PM - 18:30 PM). Furthermore, the results indicate that there is a significant improvement of bus delay in the area of CBD and eastern suburbs which can reach even a 9.5-minute delay reduction around afternoon peak hour (17:30 PM). This implies that travellers in central areas are now waiting almost 10 minutes less for their bus to arrive to the stop, which were previously delayed by severe traffic congestion most likely.

However, the area of western suburbs tends to maintain the same level of bus delay despite the new conditions and does not record a significant improvement of the bus delay reduction after the travel restrictions and stay-at-home social distancing rules. This can be explained by several factors, mostly related to the number of business hubs in that area and the type of business that are present which consist of many warehouses, bus depots, large storage areas and suburb houses. This explains that the activity in this area continues as normal and is not affected by a decrease in the number of workers travelling daily or a slight decrease in the number of personal cars on the roads.

To further illustrate the difference of delay reduction in the two selected areas, the distributions of delays from 17:30 PM to 18:30 PM are also visualised in Figure 15 in which the blue colour is used for February data set and orange colour is for March data set. Once again, it can be observed that the distribution of delays in area A shifts to the left and has a smaller tail, which confirms the overall decrease of delay in this area, while the distribution of delays in area B suffers very slight changes, remaining almost the same as previous COVID-19 travel restrictions.

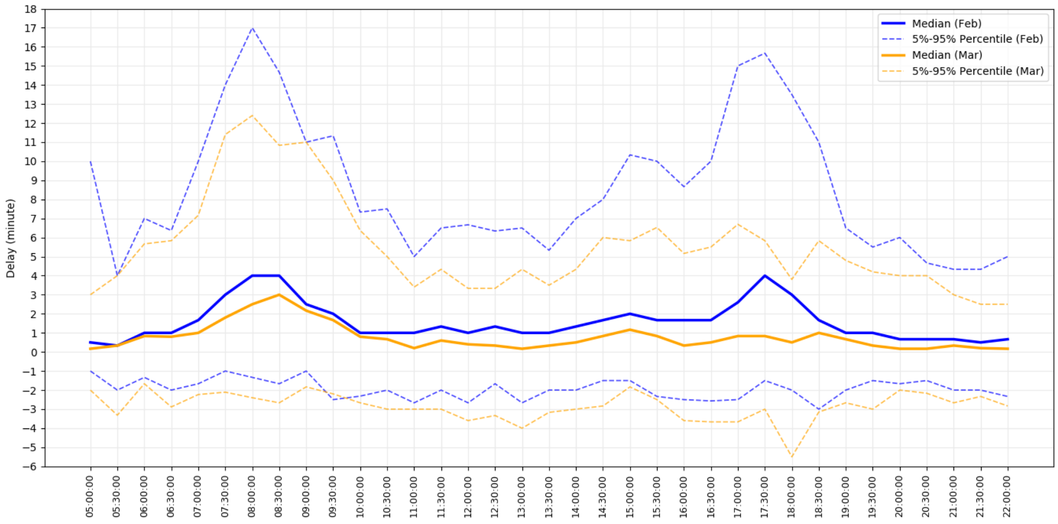


Figure 13. Bus delay in Area A. The delay significantly changed from February to March, especially the upper bound of delay in afternoon peak hours.

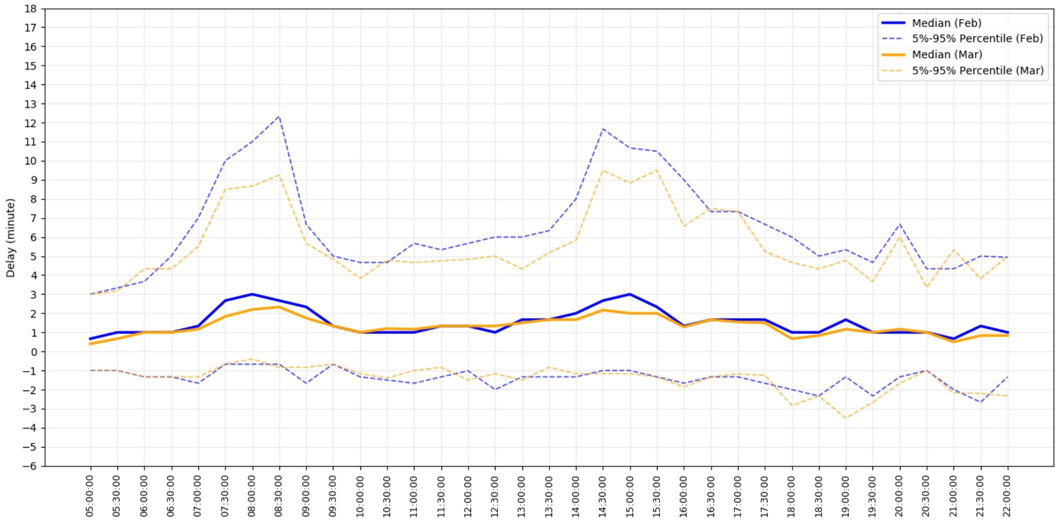


Figure 14. Bus delay in Area B. The delay slightly changed from February to March.

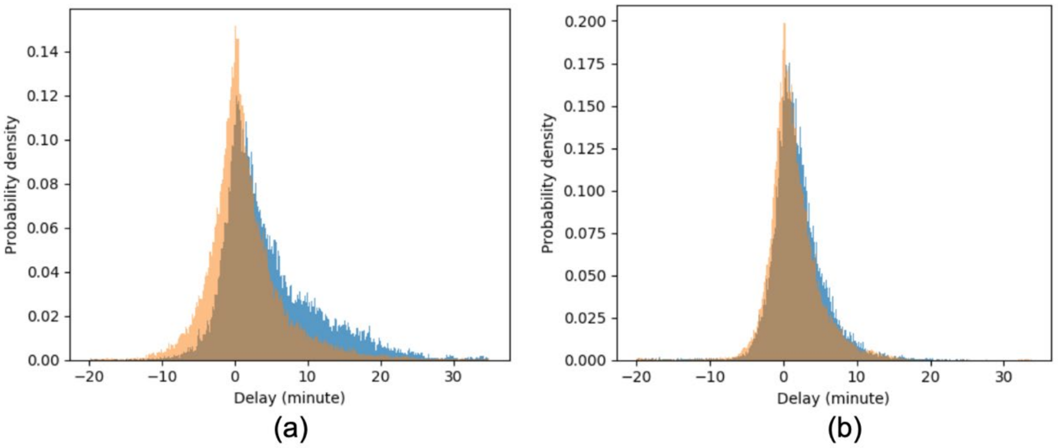


Figure 15. Distribution of delays from 17:30pm to 18:30pm in (a) area A and (b) area B. Blue is for Feburary and orange is for March 2020.

To reinforce the above findings, Figure 16 shows the delay heat map of each bus stop delay recorded for area A in morning peak hours of February and March. Each dot in Figure 16 stands for a bus stop. Green colour means smaller delay (buses arriving on time) while red colour means larger delay (buses being always late). The heat map reveals the locations where the bus delay has significantly improved, such as the central business districts, transport interchanges, shopping/entertainment centres, and beaches. Once again, this reinforces the positive impact that the travel restrictions have had in Area A and a significant improvement of the bus service in this area. The current conditions can be taken as a new ideal standard of bus operations when traffic patterns and congestion levels will be re-established to regular levels later in 2020.

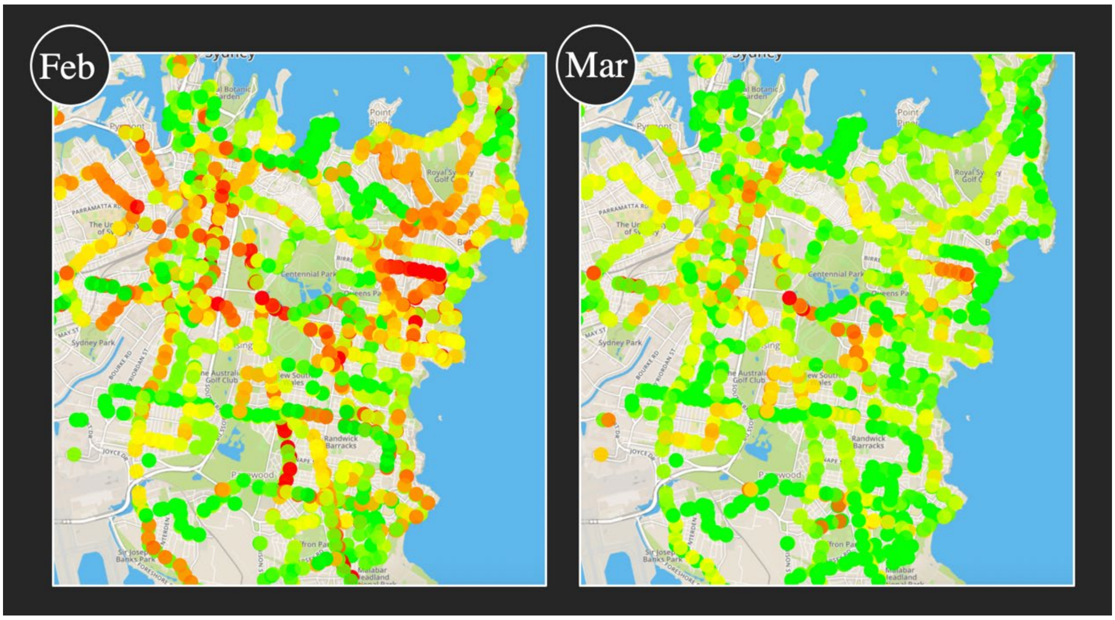


Figure 16. Bus delay heat maps for area A in morning peak hours of February and March. Read colour means larger delay.

Finally, the comparison results imply that the COVID-19 pandemic has had a different impact on public transport travel behaviour in different areas in the city, and this can be used for all bus operators making optimisation decisions in different regions across large metropolitan areas, under all variating traffic conditions: regular, pandemic, or area-restricted travel plans that the local transport agencies might impose when needing to isolate specific strategic locations in the city. The current big-data solution can be used periodically for re-assessing the above findings and for making again reinforced travel planning decisions.

5. CONCLUSION AND FUTURE WORK

In this current work we proposed a methodology to quantify the changes of bus delay, in order to study the impact of COVID-19 pandemic on public transport travel behaviours across a large metropolitan area from Australia. The research involved big-data real-time processing and detailed data analytics. The main data sets were collected from multiple sources and had various characteristics such as large-volume, real-time, spatial and temporal features. These bring true challenges to the data processing and analysis.

To verify our methodology, a case study was carried out across the Sydney metropolitan region from February to March 2020. The findings show that over the month of March 2020, the COVID-19 pandemic has significantly impacted people's travel behaviours especially in central and coastal areas. The effect of travel restrictions reduced the overall traffic congestion in the central urban areas which is usually affected by major delays on all transport modes, especially during morning/afternoon peak hours. The analytics platform revealed a significant drop in bus delays in March 2020 around the central and eastern suburbs in Sydney, reaching even a lower delay record of almost 10 minutes lower during the afternoon peak hours. This is significant improvement for bus operations in such a central location.

The proposed methodology enables us to quantify the travel behaviour changes caused by major events such as COVID-19 pandemic. This is helpful to understand and mitigate the impact in different areas with different conditions. The quantified delay reduction also reveals the potential of better transport performance, which could be used for the benchmark of transport performance improvement after the pandemic.

We keep collecting the data continuously and will extend this research by studying the travel behaviours after the pandemic, when travel restriction will be lifted. It will be interesting to see how the travel behaviours restore to the normal conditions and if the regular traffic congestion will bring the bus delays to the same levels as prior to the pandemic. A future extension would be to enable the impact of travel restrictions across multiple travel modes in the city and quantify the interchange travel time reductions. Also, we are currently exploring various mobility restriction measures that can be enforced under critical situations such as restricting completing the accessibility of travellers to specific suburbs (total lockdown of various areas). This would need to be quantified in terms of public routing as well and intermodal changes.

REFERENCES

[1] WHO. World health Organisation. 2020 April 30. Available from: URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.

[2] PM. Prime Minister of Australia. 2020 March 27. Available from URL: https://www.pm.gov.au/media/update-coronavirus-measures-270320. [Accessed 30 04 2020].

[3] Williams SD, Bock Y, Fang P, Jamason P, Nikolaidis RM, Prawirodirdjo L, Miller M, Johnson DJ. Error analysis of continuous GPS position time series. Journal of Geophysical Research: Solid Earth. 2004 Mar;109(B3).

[4] NSW, Government. Data.NSW. 2020 April 30. Available from: URL: https://data.nsw.gov.au/data/dataset/covid-19-cases-by-location/resource/21304414-1ff1-4243-a5d2-f52778048b29.

[5] Google. Google Transit APIs. 2020 April. Available from: URL: https://developers.google.com/transit

[6] Google. Public Feeds wiki. 2020 April. Available from: URL: https://code.google.com/archive/p/googletransitdatafeed/wikis/PublicFeeds.wiki

[7] Farber S, Morang MZ, Widener MJ. Temporal variability in transit-based accessibility to supermarkets. Applied Geography. 2014 Sep 1; 53:149-59.

[8] Farber S, Fu L. Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis) investments over time. Computers, Environment and Urban Systems. 2017 Mar 1; 62:30-40.

[9] Fransen K, Neutens T, Farber S, De Maeyer P, Deruyter G, Witlox F. Identifying public transport gaps using time-dependent accessibility levels. Journal of Transport Geography. 2015 Oct 1; 48:176-87.

[10] Owen A, Levinson DM. Modelling the commute mode share of transit using continuous accessibility to jobs. Transportation Research Part A: Policy and Practice. 2015 Apr 1; 74:110-22.

[11] Guthrie A, Fan Y, Das KV. Accessibility scenario analysis of a hypothetical future transit network: social equity implications of a general transit feed specification–based sketch planning tool. Transportation research record. 2017;2671(1):1-9.

[12] Hadas Y. Assessing public transport systems connectivity based on Google Transit data. Journal of Transport Geography. 2013 Dec 1; 33:105-16.

[13] Wessel N, Widener MJ. Discovering the space–time dimensions of schedule padding and delay from GTFS and real-time transit data. Journal of Geographical Systems. 2017 Jan 1;19(1):93-107.

[14] Braz T, Maciel M, Mestre DG, Andrade N, Pires CE, Queiroz AR, Santos VB. Estimating inefficiency in bus trip choices from a user perspective with schedule, positioning, and ticketing data. IEEE Transactions on Intelligent Transportation Systems. 2018 Jul 6;19(11):3630-41.

[15] Wessel N, Allen J, Farber S. Constructing a routable retrospective transit timetable from a real-time vehicle location feed and GTFS. Journal of Transport Geography. 2017 Jun 1; 62:92-7.

[16] Sun F, Pan Y, White J, Dubey A. Real-time and predictive analytics for smart public transportation decision support system. 2016 IEEE International Conference on Smart Computing (SMARTCOMP) 2016 May 18 (pp. 1-8). IEEE.

[17] Sun F, Dubey A, Samal C, Baroud H, Kulkarni C. Short-term transit decision support system using multi-task deep neural networks. 2018 IEEE International Conference on Smart Computing (SMARTCOMP) 2018 Jun 18 (pp. 155-162). IEEE.

[18] Wu J, Zhou L, Cai C, Dong F, Shen J, Sun G. Towards a General Prediction System for the Primary Delay in Urban Railways. In2019 IEEE Intelligent Transportation Systems Conference (ITSC) 2019 Oct 27 (pp. 3482-3487). IEEE.

[19] Zahabi SA, Ajzachi A, Patterson Z. Transit trip itinerary inference with GTFS and smartphone data. Transportation Research Record. 2017;2652(1):59-69.

[20] Nassir N, Khani A, Lee SG, Noh H, Hickman M. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. Transportation research record. 2011 Jan;2263(1):140-50.

[21] TfNSW, Open Data Hub. 2020 April 30. Available from: URL: https://opendata.transport.nsw.gov.au.

[22] ABS, abs.gov.au, 2020 April 30. Available from: URL: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/digital+boundaries.

[23] Hashemi M, Karimi HA. A critical review of real-time map-matching algorithms: Current issues and future directions. Computers, Environment and Urban Systems. 2014 Nov 1; 48:153-65.

[24] Lou Y, Zhang C, Zheng Y, Xie X, Wang W, Huang Y. Map-matching for low-sampling-rate GPS trajectories. in Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems 2009 Nov 4 (pp. 352-361).

[25] White CE, Bernstein D, Kornhauser AL. Some map matching algorithms for personal navigation assistants. Transportation research part c: emerging technologies. 2000 Feb 1;8(1-6):91-108.

[26] Brakatsoulas S, Pfoser D, Salas R, Wenk C. On map-matching vehicle tracking data. in Proceedings of the 31st international conference on Very large data bases 2005 Aug 30 (pp. 853-864).

[27] Srinivasan D, Cheu RL, Tan CW. Development of an improved ERP system using GPS and AI techniques. in Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems 2003 Oct 12 (Vol. 1, pp. 554-559). IEEE.

[28] Li L, Quddus M, Zhao L. High accuracy tightly coupled integrity monitoring algorithm for map-matching. Transportation Research Part C: Emerging Technologies. 2013 Nov 1; 36:13-26.

[29] Ren M, Karimi HA. Movement pattern recognition assisted map matching for pedestrian/wheelchair navigation. The journal of navigation. 2012 Oct;65(4):617-33.

[30] Varga A, Moore RK. Hidden Markov model decomposition of speech and noise. In International Conference on Acoustics, Speech, and Signal Processing 1990 Apr 3 (pp. 845-848). IEEE.

[31] Fine S, Singer Y, Tishby N. The hierarchical hidden Markov model: Analysis and applications. Machine learning. 1998 Jul 1;32(1):41-62.

[32] Forney GD. The Viterbi algorithm. Proceedings of the IEEE. 1973 Mar;61(3):268-78.